Stat 602 Project Report

Kiegan Rice and Nate Garton

Due ???

Introduction

The goal of this competition was to try to train an effective binary classifier on a small training set with many predictor variables. The criteria used to determine the best solution was the area under the ROC curve generated with predictions on the test set. The training data set had dimensions 250×300, and the test data had dimensions 19750×300.

Exploratory Data Analysis

Data Quality

The quality of the data was pristine. There were no missing values in the test or training data. There was some class imbalance in the training set (roughly 64% of the observations were class 1). A 95%, two-sided confidence interval based on large sample normality assumptions for the population class proportion was $(0.58, 0.70)$. Thus, assuming that we were truly given a random sample from the total data, it seemed unlikely that the classes were truly balanced. However, even at the upper bound of the confidence interval, the class imbalance would not likely be large enough to suggest problems with typical classification models/algorithms.

Class Conditional Distributions

Because there are so many predictor variables, it is hard to pull useful information from tables of numerical summaries. Figure 1 presents histograms of four summary statistics: minimums, maximums, means, and standard deviations for each column conditional on the class. We can see that the ranges of all of the columns are similar between and within each class. We can also see that columns means for both classes are centered at zero and look fairly Gaussian. There may be one or two "outlier" means in class 0 that are unusually small, but given the number of columns in the data, it was not immediately clear to us that this wasn't simply an artifact of randomly sampled data. Standard deviations for both classes are centered at 1, and most standard deviations for both classes are between 0.9 and 1.1.

The behavior of these summary statistics seems consistent with data where most columns are $N(0, 1)$. More here about testing normality... look at correlations... look at t-tests between column means

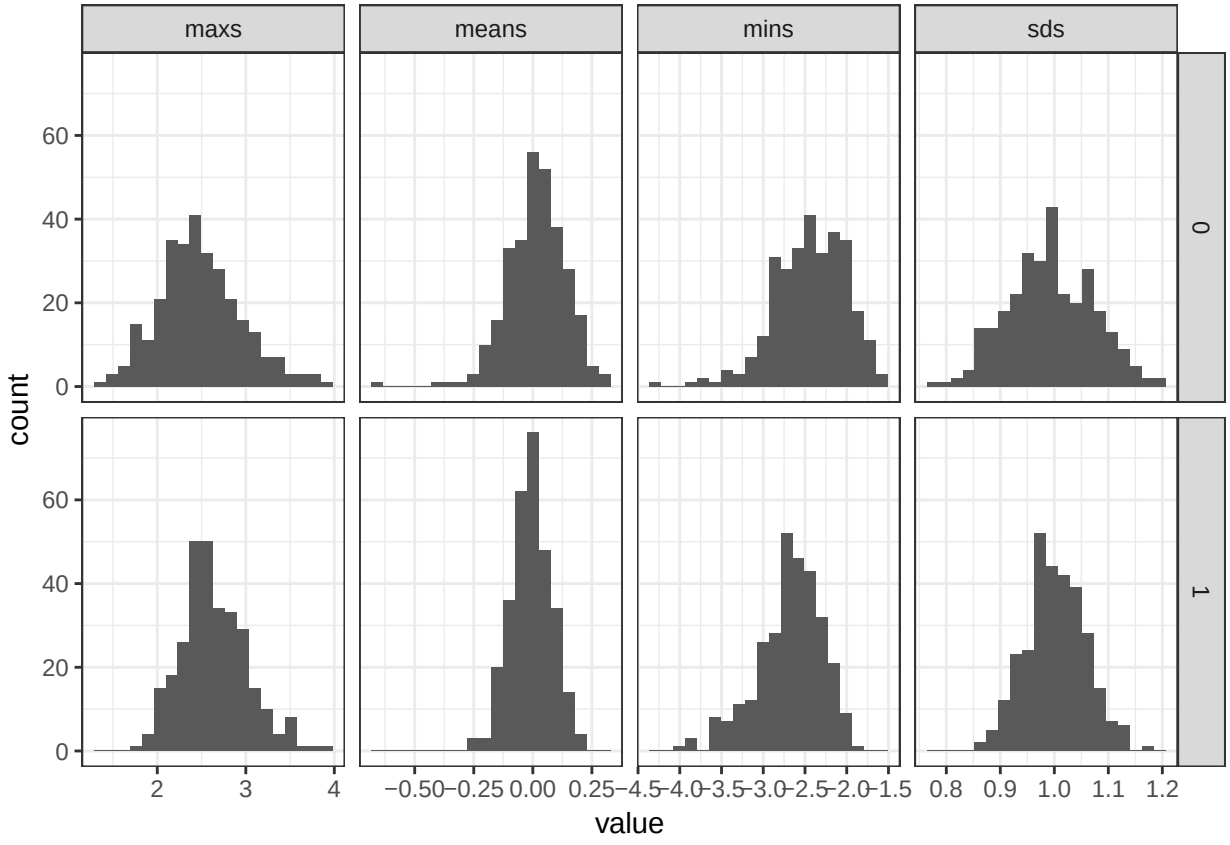

Figure 1: Class conditional histograms for column minimums, maximums, means, and standard deviations. Each row contains the column distribution for these four summary statistics for one of the two classes.